Literary Style Transfer with Sequence Models and Rejection-Pass Filtering

Daniel Kharitonov
dkh@cs.stanford.edu

Yu Fan
yuf@stanford.edu

Guneet Kohli
gkohli@stanford.edu

Abstract

In this paper we demonstrate a novel approach to literary style transfer in the natural language. We develop a new rejection-pass filter architecture for reconstructing the parallel representation of input text using a sequence-to-sequence deep network generator constrained for embeddings of samples, and evaluate the quality of text transformations for naturalness and a style score.

1 Introduction to literary style transfer

The problem of general style transfer has been extensively researched in the last ten years and led to impressive results in applying stylistic cues to still and moving images where content (contours and shapes) is easily separable from style (colors, strokes and spatial transformations). Natural language does not exhibit such separability, and therefore progress in text style transfer was hitherto limited to domains where the parallel text representations already existed.

One example of an environment with parallel representation is translation to foreign languages, where extensive corpus of references is made possible by the work of human interpreters [1]. Another example is phrase formality, where a training set encodes the formal/informal sentence pairs [2]. A third example is modern $\leftrightarrow$ archaic text adaptation, where were the stylized output can be inferred from the specialized dictionaries [3].

Literary style transfer in natural language without the parallel representations, however, so far remained an unsolved problem. Most researchers working in this field currently concern themselves with more narrow tasks – such as sentiment shifts [4], transformation of hand-coded attributes (gender, political affiliation and image captions) [5], or puzzle deciphering [6].

In this paper we permit ourselves to explore a larger goal and study the effect of literary style transfer in arbitrary texts – that is, the transformation of content from author X (content donor) into a writing style of author Y (style donor). We demonstrate the possibility of such transfer using a state of the art sequence-to-sequence generator seeded with context and constrained for specific output target, and establish a new baseline for further efforts in this area.

1.1 Main idea

The primary idea behind the rejection-pass filtering is to harness the power of generative sequence models fine-tuned for the particular literary style and conditioned on the donor content. The content seed creates the 'context' in which a generative model produces sampled sequences. Depending on how "close" this context is to the environment learned in the style training, the generative model may be able to produce candidate expansions that would be match the content donor in meaning (embeddings) while retaining the literary style acquired in training.

CS230 (Fall 2019), Stanford, CA.

For example, our content donor may include a following phrase (Ayn Rand, "Atlas Shrugged").

```
She sat at the window of the train, her head thrown back
```

If we use the first part (**"She sat at the window of the train"**) as a seed, it may activate a variety of contexts in sequence generator, which in turn will be able to generate many candidate continuations in the style it was trained in, such as: *"for half a minute, and as she looked up"*, *"her hands behind her back"*, *"artfully bent over the seat"*, and so on.

These continuations probabilistically reflect the conditional expansions that can be inferred from the combination of training and seed, and may appear highly variable (given that training corpus was large enough). If we sample several candidate continuations, we might be able to find some that will fit the meaning of a subphrase we are trying to replace – *"her head thrown back"* – with an equivalent that would be coming from a style distribution we want to enforce. Following this expansion, we can extend the seed with the phrase we just found, and repeat the entire process again. In our contrived example, the samples were not particularly creative; however, they still came from a distribution that matches the style we want to impose.

1.2 Literary Style

In this paper we will be focusing mostly on examples drawn from fiction, so we need to define the literary style (further called "style") more formally. For purposes of our study, we shall consider "style" to be the same as the "writing style", with a colloquial meaning of "being the essential elements of spelling, grammar, and punctuation, including the choice of words, sentence structure, and paragraph structure". It is important to note that while style definition is rooted in personality and language choices of a human writer, for as long as they can be discerned by an expert.

1.3 Content

Style transfer involves grafting the style from one source onto the content (story) from another. Therefore, a style transfer application features two inputs and one output, and is expected to produce the text that remains clear and approachable for an average human. It is imperative to observe that the same general content can be expressed with a large variety of linguistic devices (modifiers, metaphors, sentence formations and so on), so the total number of admissible style modifications can be very large. The degree of permissible divergence should be a tunable parameter – we would reasonably expect a different literary style to use an alternative vocabulary and idioms, but not drift too far.

2 Rejection-Pass filtering (RPF) Architecture

Our architecture relies on the ability of generative models to produce samples from the training dataset with style z and content c via latent generative distribution: $\hat{x} \sim G(z, c) = \prod p(x_t|x_{t-1..0}, c)$

With generative models, we can formulate the problem of literary text transformation for source s as finding such sequence x_i that embedding $e(s) - e(c|x_{t-1..0}) \leq \tau$. More specifically, since a trained generative model already includes style z, we can control for the embedding score by means of repeated sampling admitted on the condition of satisfying some threshold τ (a hyper-parameter).

This process of threshold-based filtering is recurring and self-adjusting: for as long as the generative model produces the output consistent with embeddings of input lexemes, the samples will be admitted; a failure to match the embedding score threshold forces the output filter to accept a lexeme from the donor text unchanged (identity transformation) and resets the context for generator. This way, the growing distances between the embedding vectors of content in donor text and generator output are periodically realigned. A complete diagram of a rejection-pass filter system is shown in Fig. 1

Our RPF implementation is a variant of a constraint-based generator conditioned on the previously transformed sequence. It bootstraps from forcing the generator to produce a sample that would randomly match the embedding score of the first lexeme; the assumption is that a well-trained generator will have enough contexts to choose from. Subsequent samples are conditioned on output already produced; the samples are accepted or rejected based on how close they are to the input lexeme that is being replaced. Two separate discriminator neural networks are employed to evaluate the metric components and adjust the generator's hyper-parameters.

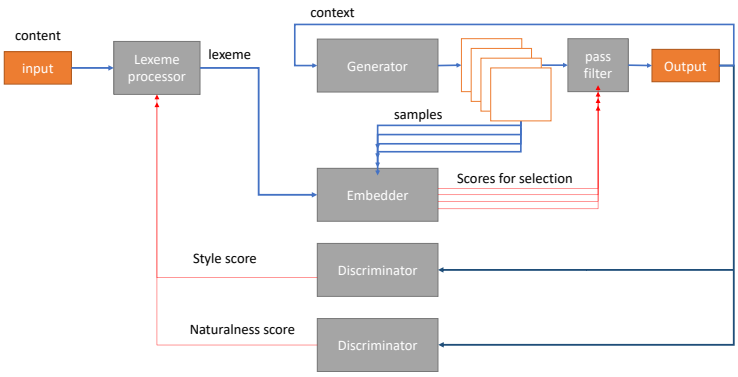


Figure 1: Reconstruction of parallel representation with rejection-pass filtering.

2.1 Solution components

Sequence-to-sequence generator network: Our algorithm relies on transfer learning: pre-trained sequence-to-sequence text generator fine-tuned on the style source. We employ an Open AI GPT-2 model with 345M parameters [9].

Lexeme processor: Takes input text and separates it into segments suitable for phrase embedder. One input lexeme can result in multiple lexemes spun by the generator; such behavior permits the style transfer engine to replicate the sentence structure and punctuation of the style donor. In our model we are using a relatively simple heuristics to determine the sentence features. The trade-off between the long and short lexemes reflects our preferences for precision of replacements.

Embedder Phrase embedder calculates similarity scores for the content input lexeme and the candidate expansions (samples) produced by the generator. Encoder precision significantly depends on the length of the input lexemes; longer phrases are more difficult to encode and stand a lower chance to be matched against a sample from the generator. We employ Google Sentence Encoder to calculate the candidate lexeme scores and compare them to content input.

Pass filter Pass filter is a simple logical gate admitting the best sample that clears the embedding threshold. If the sample set fails to clear the filter, the input lexeme is admitted unchanged.

Discriminators The consensus of the research community on machine-generated text appears to be a score that includes triple measures measures style, narrative fluency, and content equivalence (e.g. Mir et. al [10]). We further consider that the two dimensions of the metric (naturalness and content preservation) are of the satisficing type, while "style" is the metric component we optimize for. Insofar, all three components appear insufficiently defined in the research literature.

We compute "style transfer score" as a shift in a likelihood of cross-style attribution (difference in P(style author|text) before and after the transformation) by a DistilBERT network. In other words, we train a style discriminator to produce softmax probabilities for recognizing the author, with an expectation that generator output should look "less" like the corpus of content author and "more" like the corpus of style donor. A positive difference in attribution probabilities indicates success of the style transfer [1].

For "naturalness score", we looked at state-of-the-art fake text detection systems like GLTR and GROVER [11, 12], but found them poor in precision because our content (fiction by human authors) was frequently labeled "fake" in tests. For lack of existing naturalness detectors trained on fiction, we decided to calculate our own with an LSTM neural network adversarially trained to recognize machine-generated text. After the error analysis we found our naturalness classifier to have a low recall rate (predicting most generated samples to be highly natural). Therefore, we finally chose to employ the AWS Mechanical Turk workers to compute our naturalness metrics. We provided the randomly shuffled sample of 60 style-transformed texts supplemented with 140 original author texts for evaluation and computed the scores by averaging the responses [2].

[1] see Appendix B.3 for more details of our style classifier

[2] see Appendix B.4 for more details of our LSTM-based classifier attempt

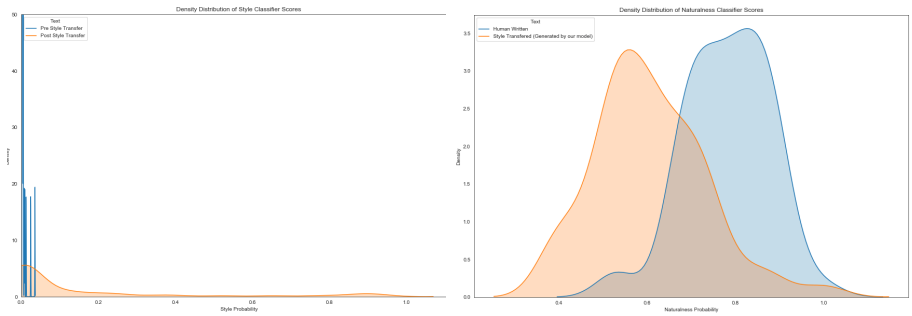


Figure 2: Naturalness distributions shift after RPF (blue: before, red: after)

Table 1: Acceptance rate for replacement lexemes in content coming from different authors

	Twain	Hemingway	Shakespeare	Dumas	Austen
Percentage replaced	0.212	0.41	0.263	0.421	0.473

We chose not to implement a separate content discriminator, as the content score can be calculated directly in the generator when producing the output.

3 Dataset and the generator hyperparameters

To train and evaluate our initial model, we created three distinct style datasets compiled from texts in the Project Gutenberg online library. Specifically, we are using the corpora of texts by Alexander Dumas (8.3MB), Jane Austen (4.1MB) and Vladimir Nabokov (9.8MB).

We cleaned these datasets from the publishing notes and annotations to train three sequence-to-sequence style generator instances (by means of fine-tuning the pretrained GPT-2 model). In addition, we sliced these texts into 600B chunks, cleaned them off proper and personal names using a small named entity removal (NER network), and labeled chunks with author codes to train our style discriminator network. The training sets were balanced for presence of snippets by every author. We used discriminator outputs to search the hyperparameter for the generator; the list of the final hyperparameters values is provided in Appendix A.

4 Experimental results

We have employed our RPF system to generate approximately 100 text excerpts in three different styles (Nabokov, Dumas, Austen) [3]. We have observed an average change in the probability of style cross-attribution of +14.4%, and an average change in the text naturalness of -45% percent (Fleiss Kappa = 0.08). In other words, RPF achieved a noticeable increase in style attribution probability, at cost of dropping for naturalness. The distribution of style deltas, however, is far from the uniform. In Fig. 2 we can see that while some excerpts were transformed well enough to fool the style discriminator, most stylistic changes were relatively subtle.

To check for a possibility of trading style for naturalness, we plotted a joint distribution of style and naturalness changes in our samples (Fig. 3). Judging from that plot, it seems like the loss of naturalness is fairly uniform across the samples.

4.1 Discussion and Error analysis

As shown in the graphs, RPF system works as designed, and is able to shift the donor style effectively. One interesting statistics is the ability of RPF to work across different genres. In the table below, five different content sources are compared for lexeme replacement rates in the writing style of Nabokov:

[3] some transformation examples are provided in Appendix C

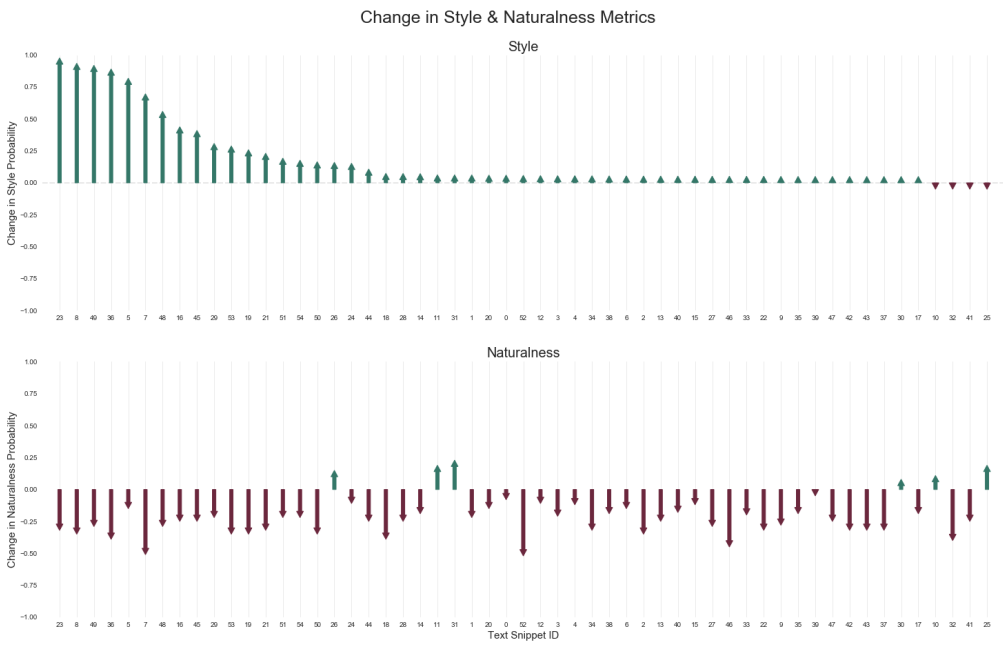


Figure 3: Probability of cross-attribution after RPF: style (top), naturalness (bottom)

Somewhat unsurprisingly, we observe that the donor content has higher chances to be transformed if it was reasonably close to the style to begin with. The lowest transformation score is of Mark Twain ("The Adventures of Huckleberry Finn") who employs many short phrases written from the first-person viewpoint; Nabokov style appears hard to graft on such text. Shakespeare is a close second on hardness of transformation, which means his archaic language does not stimulate the contexts of Nabokov-trained generator effectively [4].

For error analysis of RPF, we hand-checked the outputs that appeared highly unnatural, and binned the issues into five categories: imprecise replacements (embedder matched the original content poorly), broken sentence structures (heuristics did not work for punctuation), content leakage from style donor (unwanted names or terms from the style corpus), unsatisfied long-range dependencies (mismatch between the seed and replacement content over features described earlier), unintended overtones (style-induced biases on original content).

Conclusions and future work

This paper outlines an original rejection-pass filtering algorithm built alongside the three-component evaluation metric, which sets a new baseline for literary style transfer. The idea of RPF is straightforward, modular and can be readily improved once better sequence generation and embedding models become available.

Owing to a relative novelty of an application, we set a new benchmark for the field, and demonstrate the use of the scoring metrics to gauge the performance of style transfer in literary texts. The future work can focus on improving the treatment of names and the long-range dependencies.

We think that some of the issues we discovered can be addressed in the future. Most importantly, a better sequence generator and sentence encoder may alleviate content mismatches and long-range dependence issues. Leakage of proper names appears a harder issue and may require a separate framework for alleviation. Finally, we thing that style-induced bias is an irreducible type of error because it properly reflects the attitude of style donor towards a familiar context.

[4] More detailed information on performance of generator and discriminators is provided in Appendix B

Code

The code for this project is available online and can be reproduced on a machine with NVIDIA V100 (or better) GPU: https://github.com/volkfox/text_style_transfer

Acknowledgments

The authors owe a lot of debt to many researchers contributing to computational natural language field in the last decade. This project would not be possible without Google Colab, Amazon AWS, and OpenAI GPT-2 models.

References

[1] Orch, Franz (April 28, 2006). "Statistical machine translation live". Google Research Blog. Google.

[2] Rao, S., Tetreault J. (2018) Dear Sir or Madam, May I introduce the YAFC Corpus: Corpus, Benchmarks and Metrics for Formality Style Transfer, NAACL-HLT

[3] Jhamtani H., Gangal v., Nyberg E. "Shakespearizing Modern Language Using Copy-Enriched Sequence-to-Sequence Models"

[4] John V., Mou L., Bahuleyan H., Vechtomova O. "Disentangled Representation Learning for Non-Parallel Text Style Transfer"

[5] Sudhakar A., Upadhyay B., Maheswaran A. "Transforming Delete, Retrieve, Generate Approach for Controlled Text Style Transfer"

[6] Lample G., Subramanian S., Denoyer L. "Multi-attribute Text Rewriting"

[7] Yang Z., Hu Z., Dyer C., Xing E., Berg-Kirkpatrick T. "Unsupervised Text Style Transfer using Language Models as Discriminators"

[8] Cera D., Yinfei Y., Sneg0yi K., Nan H., Nicole L., Rhomni J., Noah C., Marco G., Steve Y., Chrs T., Yun-Hsuan S., Brian S., Ray K., "Universal Sentence Encoder", Google Research 2018

[9] Open AI GPT-2. https://github.com/openai/gpt-2

[10] Mir R., Felbo B., Obradovich N., Rahwan I. "Evaluating Style Transfer for Text" https://arxiv.org/abs/1904.02295

[11] Catching a Unicorn with GLTR: A tool to detect automatically generated text http://gltr.io/dist/index.html

[12] GROVER: A State-of-the-Art Defense against Neural Fake News. https://grover.allenai.org/

[13] Finkel J., Grenager T., and Manning C., 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)

[14] Generating Sentences from a Continuous Space. arXiv preprint arXiv:1511.06349v4

[15] Kiros Y., Zhu Y., Salakhutdinov R., et al. 2015. Skip-thought vectors. arXiv preprint arXiv:1506.06726

[16] Sanh, V., Debut, L., Chaumond, J. and Wolf, T., 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

Appendices

A Generator Hyperparameter keys

Table 2: Generator hyperparameters

Rejection threshold	minwords	maxwords	seed	nsamples	temperature
0.68	2	7	10	10,000	1.0

Rejection threshold – defines the minimum semantic similarity between phrase embeddings from the content input and samples the filter could accept. Increasing this parameters would force the model to search for better matches at risk of not finding the good candidates.

Minwords, maxwords – are the hyperparameters for lexeme processor that define the minimum and maximum number words in the next phrase for processing. A lexeme preferentially ends with a natural separator (period, comma, or semicolon), but a long input phrase may require an early stop.

seed – is the size of the seed phrase (context length measured in lexemes) given to the generator. More lexemes in the seed will result in the deeper context and may produce better samples, but can also restrict the generator to a cue that appears too narrow.

nsamples – the number of samples the generator produces per every new context. Larger numbers allow more candidates for rejection filter to choose from, but require more processing time to generate them and calculate embedding scores for them.

temperature – the generative network hyperparameter that describes a relative likelihood of produced samples. Higher values of temperature produces less likely – i.e. "more creative" – expansions.

B Performance and architecture insights

B.1 Generator performance statistics

Generator performance statistics relates to ability of producing content expansions that can pass the rejection threshold (measure of embedding similarity to lexemes being replaced). In our experiments, this ability appears limited by two factors: (1) the amount of fine-tuning the GPT-2 model gets over style corpus and (2) the "closeness" of content donor to the contexts of target style.

To study the effect of fine-tuning, we kept statistics for the number of successful replacements per 20 attempts for various number of training cycles. Somewhat counter-intuitively, the percentage of lexemes successfully replaced quickly drops along with the fine-tuning; while it typically starts with about 70% without style-specific training, it goes down to 40-50% once fine-tuning is complete (see Fig. 4).

This result is a bit unexpected and suggests that the GPT-2 model is originally trained on the Internet data providing a broad variety of contexts; fine-tuning to the specific style corpus "narrows down" this model. It is an interesting finding as it tells us something about the inner structure of generative NLP models. Another observation generator worth mentioning is that is frequently capable of matching the content lexeme verbatim with stock model; this almost never happens after fine-tuning and again implies that the "general knowledge" of language gained over the Internet databases in GPT-2 is more variable than any particular style commanded by a specific writer.

A different way to look at performance is to check the number of permissible expansions per sampling batch. That is, conditioned on the fact that we found a replacement, how many other samples could have cleared the threshold?

Somewhat surprisingly, and given the 10,000 samples we draw from the generator per input, we observe that the majority of contexts where we do find a replacement actually feature ten or less feasible candidates (see Fig. 5)

This implies the high specificity of context switching based on seed: it is reasonable to assume that style corpus includes relatively few excerpts that would match an arbitrary seed from a different text; a small expected number of candidate replacements means the network has to work extra hard to produce continuations (in practice, we can address this problem with increasing the number of samples, or the generation temperature).

Next, we can get some insights on how RPF works by looking at probability of generating a successful replacement conditioned on the prior event. For example, in a generator trained on Alexander Dumas style and transforming a text by Ayn Rand, the average probability of successfully replacing a random lexeme is 48%, but

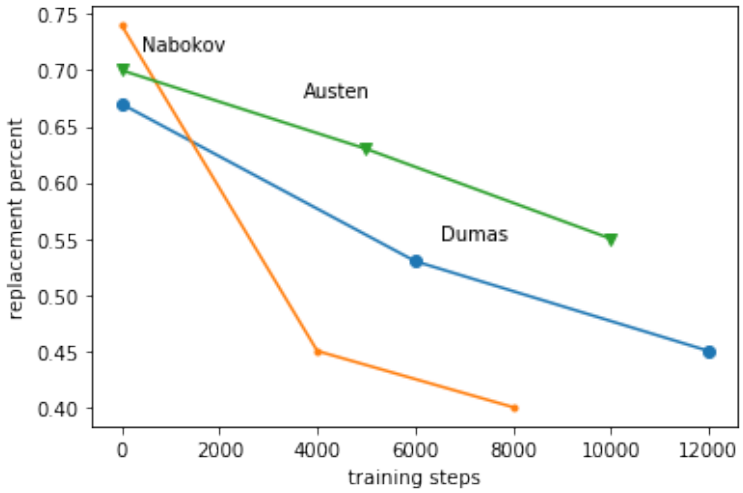


Figure 4: Reduction in GPT-2 lexeme sampling performance after fine-tuning for three author styles.

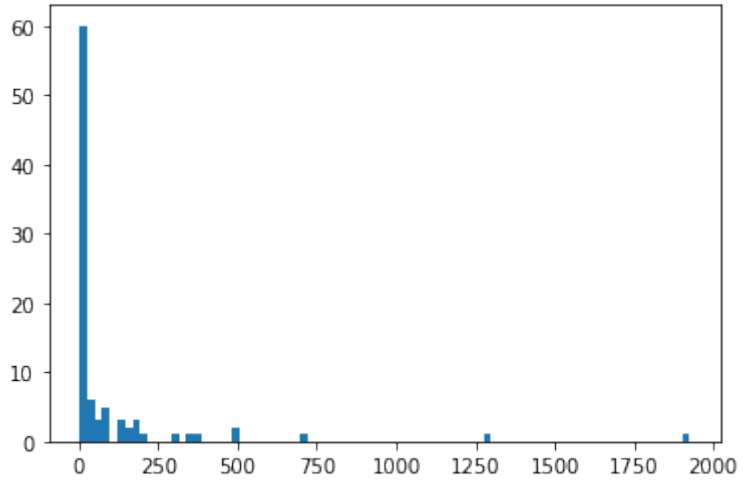


Figure 5: Histogram of acceptable candidates per successful replacement.

replacement events do not appear to be independently distributed. Instead, the observed probability of replacing a next lexeme given the previous one was replaced is $p(success|success) = 0.23$, while a probability of replacing a next lexeme if the previous lexeme was not replaced is much higher, at $p(success|failure) = 0.72$.

Such asymmetry can be explained with the logic of conditional generation: once lexeme was replaced, the seed now includes the context more familiar to the generator, and the output becomes more specific; however, the resulting set of extension candidates may not fit the original (content donor) story anymore, so all samples get rejected and the output is resynced by means of copying the next lexeme from the content source.

B.2 Style and Naturalness Discriminators

Since our intention is to provide machine-based style transfer, we need to task ourselves with subject of output evaluation. It follows from our discussion on styling that such score should include the measures of a triplet style source, narrative fluency, and content equivalence. Provided that our ultimate goal is a perfect imitation of source style conditioned on story from content source, missing either of the aforementioned factors will not yield a satisfactory result. For one example, if the output text does not employ the vocabulary and sentence structure of style donor, it will result in the stylistic miss. For another example, if the output employs the style but departs from the content, it will fail to form a parallel representation. For a third example, if the output text successfully fuses the content with style of input sources but violates general language and writing norms, it will

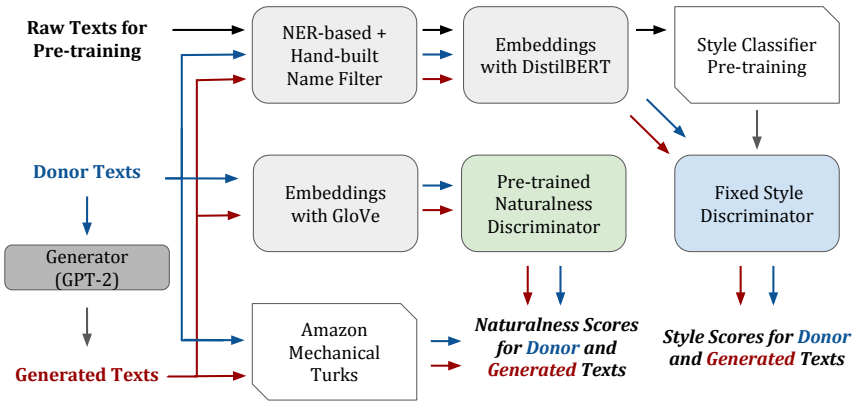


Figure 6: Architecture and training pipeline for style and naturalness discriminators.

result in a poor reading experience. Therefore, to evaluate the quality of style transfer, we need to take all those considerations into account.

In evaluating results of the literature style transfer, we must consider that the two dimensions of the metric (naturalness and content preservation) are of the satisficing type, while the style is the metric component we optimize for.

Before discussing the architecture of style and naturalness discriminators, we need to establish their starting points and limits. For the style discriminator, the starting point (default) is fairly intuitive and consists of the initial cross-authorship attribution probability measured on the content donor (the untransformed text). That is, a change in style is defined as a shift in the probability of attribution towards the author of donor style. For example, a text from Ayn Rand can have the initial cross-attribution score to Nabokov of P(Nabokov | text by Ayn Rand) = 0.32, so ideally we would like a transformed text to have P(Nabokov | text by Ayn Rand modified by RPF) = 1.0 (upper limit of measure), for a change of +0.68.

The case for naturalness discriminator is a bit more complicated because the ground truth is based on the subjective judgement of humans about this particular piece of text. Our ultimate goal should be able to recognize every RPF-created piece as human-written, but the starting point can be variable because some fiction literature may appear mistaken for machine-generated text. As in the case of the style classifier, we intend to collect the difference between the initial probability of misattribution P("fake" | original content), and the same probability after the transformation: P("fake" | transformed content).

The overall architecture for training the style and naturalness discriminators is shown in Fig. 6.

B.3 Style Discriminator Architecture Selection and Performance

To get a feeling for style baseline, we started with a simple "bag of words" (count vectorizer) architecture for attribution. This gave us an early insight that the easiest way to attribute a piece to an author is to encounter proper names characteristic for this writer. For example, in Fig. 7 we can learn that Alexandre Dumas Jr. frequently mentions names like Porthos, Artagnan and Athos and never mentions Fanny, Emma or Elinor (the latter would be characteristic to Austen Rand). Such observation makes a lot of intuitive sense: if you are given an excerpt by an unknown author and catch a glimpse of the word "Frankenstein" you will be reasonably sure you are looking at work by the Mary Shelley even if you never have read her book.

This efficiency of the proper names for attribution, however, creates a problem if we are trying to detect style shift from one writer to another: since the proper names in the story arc should not change, alterations in use of metaphors, synonyms, or punctuation applied by the RPF can be deemed insignificant by a style classifier.

Therefore, our next step was to pre-process the texts by removing named entities to shift the attention of classifier to language and sentence structures rather than names. To achieve that, we designed a two-step filter: first, we used the pre-trained Stanford's Named Entity Recognizer [13] to identify all tokens with label <PERSON>, and second, we train a logistic regression classifier on a bag of words feature set and rank every word/token by their

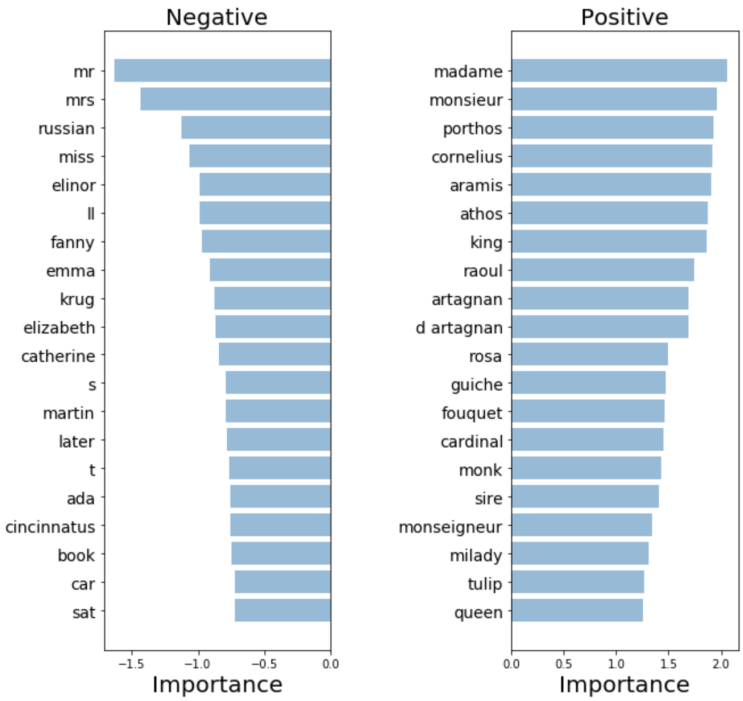


Figure 7: Top Words Associated with Alexandre Dumas Jr.

Features	**Logistic Regression**	**Xgboost**	**SVM**	**Multinomial NB**	**DNN**	**GRU**	**LSTM**
Count Vectorizer	**0.13 / 0.96**	-	-	0.18 / 0.97	-	-	-
TF-IDF	0.33 / 0.96	-	-	**0.23 / 0.98**	-	-	-
TF-IDF + SVD	-	0.21 / 0.92	**0.16 / 0.95**	-	-	-	-
Glove	0.24 / 0.96	-	-	-	0.16/ 0.96	**0.13 / 0.96**	0.20 / 0.96
DistilBERT					**0.04 / 0.99**		

Figure 8: Model Comparisons for Texts with Name Masking (Score: Log-loss / Macro-F1 Score)

corresponding coefficient obtained from the model. Accordingly, we prepared a masking set which contains top-50 tokens with the highest coefficients. We apply this mask to the dataset used for style attribution to ensure that highly influential rare names were masked.

After setting the aforementioned name filters, we ended up with 9,000 pieces of pre-processed texts, and 3,000 pieces for each author, respectively. Furthermore, we divide them into three datasets with 80/10/10 percent for training, developing as well as testing the models. After that, we applied several well-known classifier architectures to see which perform the best (see Fig. 8). We have also kept their performance without mask (see Fig. 9) to check how much information was lost by removing named entities.

We found that the best classification results are achieved with a pre-trained DistilBERT model, proposed by [16]. This was in line with out expectations, as BERT should be able to pick on long-term features like sentence structure and punctuation that would go unnnoticed with simpler classifiers. We utilized BERT with a fully connected feed-forward network with dense skip-connections on top to perform the discrimination. The downstream classifier entails 6 hidden layers and around 29M parameters. The optimal hyper-parameters were identified by randomly searching through learning rates, sizes of each layer and mini-batch sizes.

Features	Logistic Regression	Xgboost	SVM	Multinomial NB	DNN	GRU	LSTM
Count Vectorizer	**0.04 / 0.99**	-	-	**0.04 / 0.99**	-	-	-
TF-IDF	0.15 / 0.98	-	-	**0.08 / 0.98**	-	-	-
TF-IDF + SVD	-	**0.04 / 0.99**	0.05 / 0.98	-	-	-	-
Glove	0.20 / 0.96	-	-	-	0.08/ 0.98	**0.06 / 0.98**	0.07 / 0.98
DistilBERT					**0.01 / 1.00**		

Figure 9: Model Comparisons for Texts without Name Masking (Score: Log-loss / Macro-F1 Score)

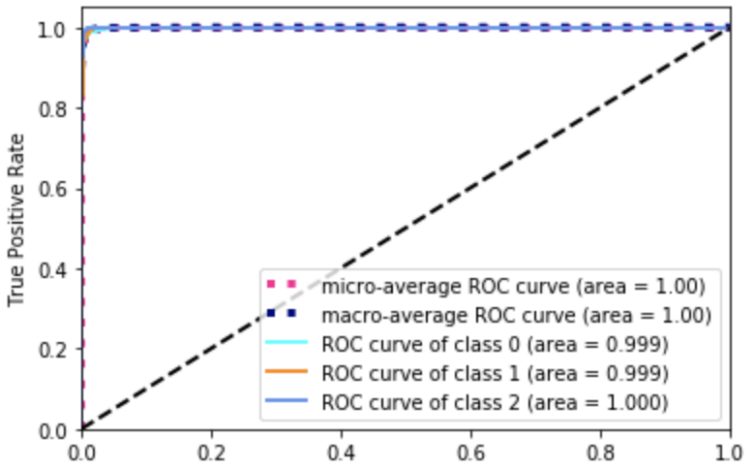


Figure 10: Evaluation of Pre-trained Style Discriminator

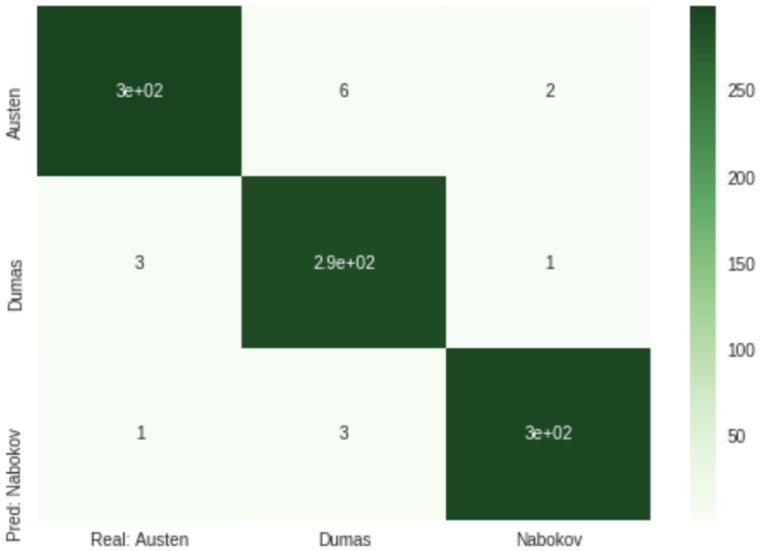


Figure 11: Confusion Matrix of the Style Discriminator

Results: As shown in Fig. 10, our style discrimination model achieves a nearly perfect performance with named entities removed, with an overall cross-entropy loss of 0.04, and a nearly 1.00 of AUC for each class when evaluated on test set. As shown as well in the confusion matrix (see Fig. 11), only few cases are misclassified

between the three authors in our corpus. This gave us good confidence that we can measure the style shifts in our transformations reliably.

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_2 (InputLayer)         (None, 30)                0
_________________________________________________________________
embedding_2 (Embedding)      (None, 30, 256)           2419456
_________________________________________________________________
lstm_2 (LSTM)                (None, 128)               197120
_________________________________________________________________
dense_2 (Dense)              (None, 1)                 129
=================================================================
Total params: 2,616,705
Trainable params: 2,616,705
Non-trainable params: 0
```

Figure 12: Architecture of Pre-trained Naturalness Discriminator

B.4 Naturalness Discriminator Architecture Selection and Performance

After faling to find a pre-trained 'fake text' classifier that would perform well on fiction literature, we have decided to leverage the adversarial classification done by Mir et. al [10] and build our own discriminator. We selected this architecture because the authors have claimed that their technique is strongly correlated and in agreement with human judgement.

The resulting naturalness discriminator is a pre-trained LSTM classifier (Fig. 12) that reads the input text via an embedding layer and produces a binary prediction after seeing the final token. This classifier has 2.6M weights. The model is trained on the Books Corpus introduced in Kiros et. al. [15] and is a collection of text from 12k e-books mostly, fiction. As per [14] the training is performed using early stopping on a 80/10/10 train/dev/split of 320k samples constructing a dataset of 50% real samples acting as positive examples and 50% imputed examples from the corpus acting as negative examples.

Results: As shown in Fig. 13, most of our samples were classified to be around zero change in naturalness, indicating that style transfers are roughly similar in term of its readability in comparison to their human-written counterparts from donor texts. Upon investigation, we realized that our naturalness classifier simply predicts most samples to be highly organic (probability > 0.9) even when they are not. We tried setting different thresholds for the decision boundary however saw no change.

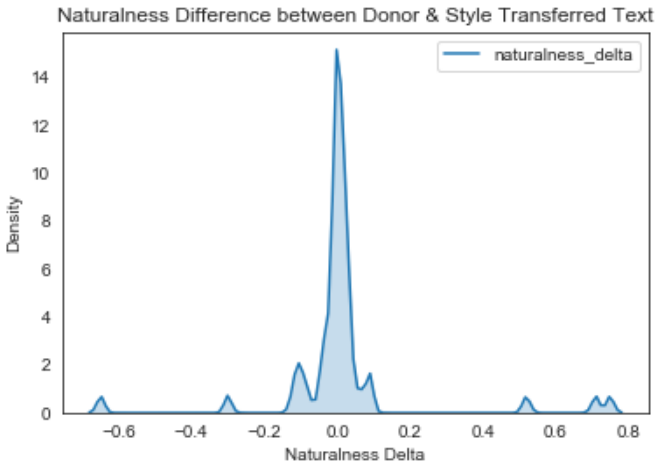


Figure 13: Distribution of Naturalness Shift between Donor and Style Transferred Texts

To check whether our samples appear indeed highly natural to the humans, we employed 'masters level certified' mechanical turkers to annotate a randomly shuffled sample of our 55 style generated texts supplemented with 140 original author texts. Annotators are instructed to mark each text snippet on a 5 point scale as follows:

- 'Completely Natural'
- 'Mostly Natural',
- 'Equally Natural & Unnatural',
- 'Mostly Unnatural'
- 'Completely Unnatural'.

Multiple workers were assigned to each sample, and we averaged their results. To exclude the inattentive workers, 5 spiked examples containing a randomly placed 'N/A' words were included in evaluation set. The annotators are instructed to select 'N/A' as a rating if they find such a spike in their annotation data. See Fig. 14 for a sample of our annotation workflow.

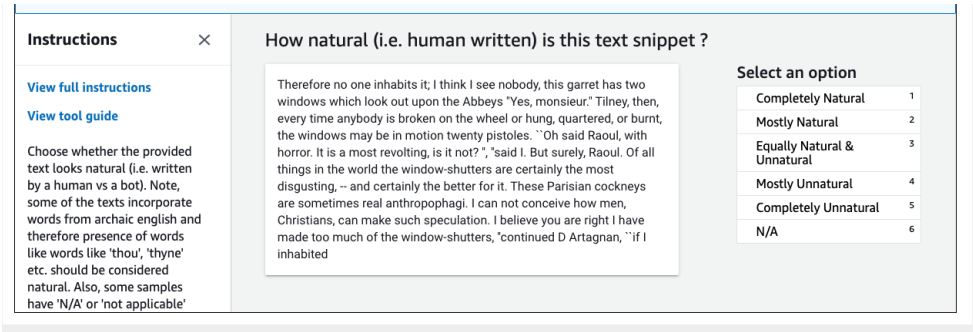


Figure 14: Mechanical Turk UI Design for Annotation Workflow

Naturalness Classifier Error Analysis

After obtaining results from the Mechanical Turk, we used them as ground truth to check our classifier. Unfortunately, the results were not satisfyig as distributions of scores did not match well (see Fig. 15). We observed that turkers rated our samples much lower than a classifier, at an average 45% loss in misattribution probability per transformation.

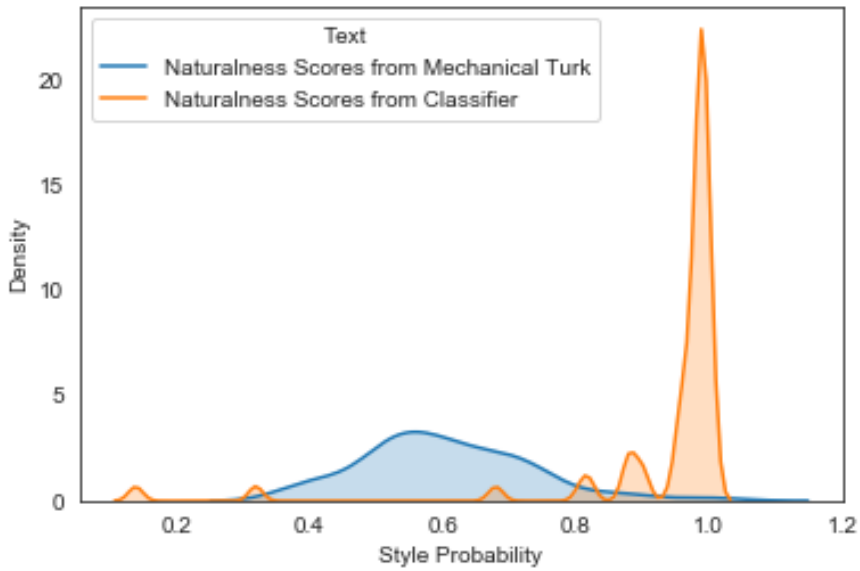


Figure 15: Distribution of scores by Mechanical Turkers vs classifier

While doing the debugging, we have found that while our classifier can pick out cases which are clearly natural or unnatural, it struggled on tasks which could be marked as 'Equally Natural Unnatural'(probability score =

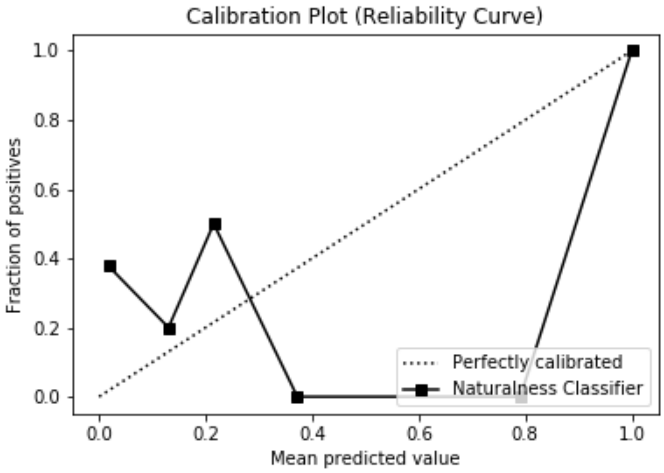


Figure 16: Calibration of Naturalness Classifier in Comparison to Mechanical Turk Scores

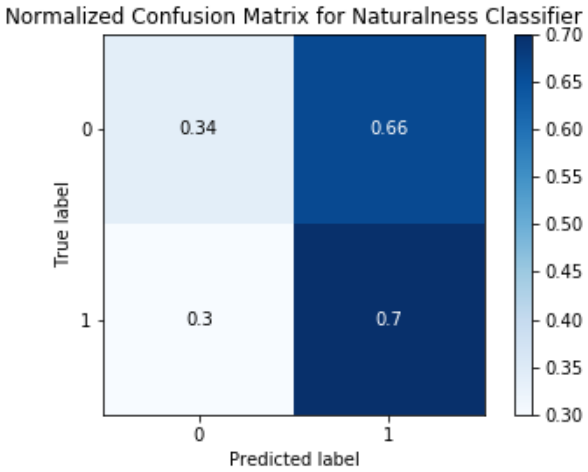
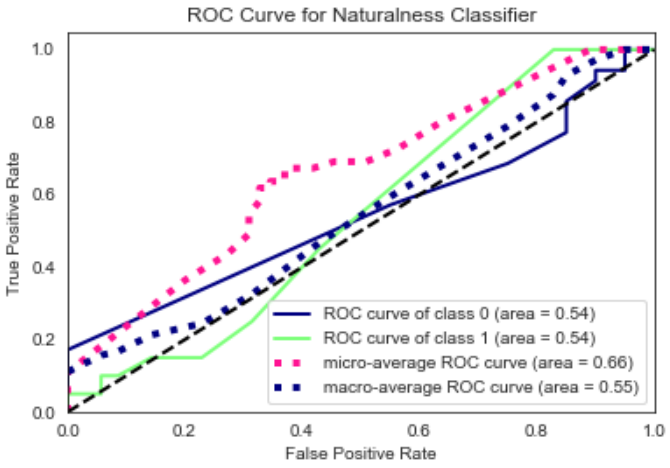


Figure 17: Confusion Matrix for Style Transferred Texts for Naturalness Classifier

0.6) and 'Mostly Unnatural' (probability score = 0.4) as evident in the calibration plot in Fig. 16. We attribute this artifact to a small sequence length (n=30) used in the pre-trained model as well as the difficulty in gauging the long-term relations in text for machines. On the opposite, we found the results from mechanical turk to be well-balanced in terms of variation of the naturalness probability scores obtained.

After lingering over those results, we have decided to use human judgements as naturalness scores in our project going forward. We are providing the confusion matrix, ROC curve and Precision Recall curves for our LSTM naturalness classifier here as references.



[h!]

Figure 18: Receiver Operator Characteristics for Style Transferred Texts

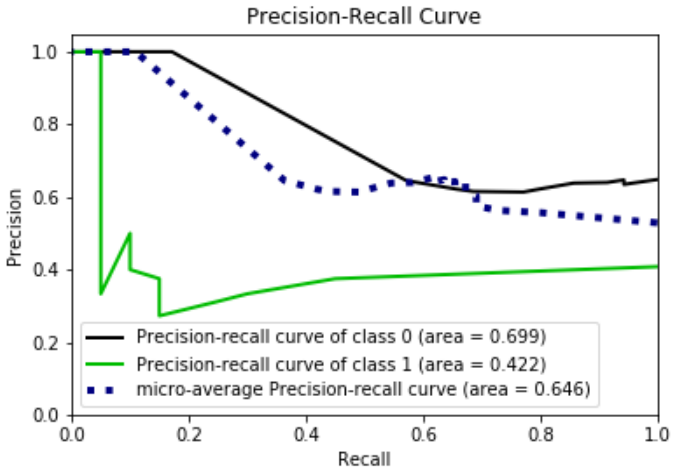


Figure 19: Precision Recall Curve for Style Transferred Texts

C Content example: Ayn Rand, "Atlas Shrugged"

She sat at the window of the train, her head thrown back, one leg stretched across to the empty seat before her. The window frame trembled with the speed of the motion, the pane hung over empty darkness, and dots of light slashed across the glass as luminous streaks, once in a while.

Her leg, sculptured by the tight sheen of the stocking, its long line running straight, over an arched instep, to the tip of a foot in a high-heeled pump, had a feminine elegance that seemed out of place in the dusty train car and oddly incongruous with the rest of her. She wore a battered camel's hair coat that had been expensive, wrapped shapelessly about her slender, nervous body. The coat collar was raised to the slanting brim of her hat. A sweep of brown hair fell back, almost touching the line of her shoulders. Her face was made of angular planes, the shape of her mouth clear-cut, a sensual mouth held closed with inflexible precision. She kept her hands in the coat pockets, her posture taut, as if she resented immobility, and unfeminine, as if she were unconscious of her own body and that it was a woman's body. She sat listening to the music. It was a symphony of triumph. The notes flowed up, they spoke of rising and they were the rising itself, they were the essence and the form of upward motion, they seemed to embody every

human act and thought that had ascent as its motive. It was a sunburst of sound, breaking out of hiding and spreading open. It had the freedom of release and the tension of purpose. It swept space clean, and left nothing but the joy of an unobstructed effort. Only a faint echo within the sounds spoke of that from which the music had escaped, but spoke in laughing astonishment at the discovery that there was no ugliness or pain, and there never had had to be. It was the song of an immense deliverance.

Style score (triangulated to Austen, Dumas, Nabokov): [0.01131631, 0.0077914 , 0.98089224]

Classifier naturalness score: 0.8814134

C.1 Machine Generated: Nabokov style transfer

She sat at the window of the train, her head thrown back, one leg stretched across to the empty seat before her. The window frame trembled with the speed of the motion, the blackness of space, and dots of light slashed across the glass as luminous streaks, once in a while. Her leg, sculptured by the tight sheen of the stocking, its long line running straight, over an arched instep, to the tip of a foot in a high-heeled pump, had a feminine elegance that seemed out of place in the dusty train car and oddly relatable in the low-slung ranks of her. She wore a battered camel's hair coat that had been expensive, wrapped shapelessly about her slender, nervous body. The coat was clean-cut too, worn into the sleek collar that clung to her neck. A sweep of brown hair fell back, down, down, over her shoulders. Her forehead, above the sun-dappled asphalt, was smudgy and plump, and the wing of her nose, a glossy dark-orange, was exactly shaped as her lips, small stern Plebeian nose thrust below, with downward curved mouth. Her long legs were free, her pink mane kept catching the rays, and the muscles of her hands continued to embrace, as if she suddenly slipped into a woman's body. The silence got broken. It was a symphony of triumph. The notes flowed up, they spoke of rising and they were the rising itself, they were the essence and the form of upward motion, they seemed to embody every human act and thought that had ascent as its motive. It was a sunburst of sound, breaking out of hiding and spreading open. It had the freedom of release and the tension of purpose.

Style score difference towards Nabokov: + 0.0032442212

Classifier naturalness score difference towards the original: + 0.05819315

C.2 Machine Generated: Jane Austen style transfer

She sat at the window of the train for half a minute, and as she looked up, she stretched one leg across to the empty seat in front of her. The window frame trembled with the speed of the motion, the unnerveness of the table amidst the scattered glasses, and the distressing tones of the glass pressed against the glass as luminous streaks obtruded the view. Her leg, sculptured by the tight sheen of the stocking, its long line running straight, from the ankle to the fore-foot, to the tip of a foot in a glide, had a feminine elegance that seemed out of place in the dusty train car and oddly incongruous with the rest of her. She was a scantily shaven woman: her hair was very short but expensively cut, her face was a little hoary and attached to a rather pensive, nervous body. The coat collar was raised to the slanting yellow humbug of her hat. A sweep of brown hair fell back, affording a quick view of her face. Her hair was striking, as if unsubdued by ill-timed gravity, and almost turned into a veil against her face, her sensual mouth held closed with inflexible precision. She put one hand tied behind her face, and her heart leaped against the stomach with sudden feeling, as the cottony, trembling hand was about to embrace body her woman's body. She was in a reverie, as he entered the room. It was a symphony of triumph. The Tilneys were at the instrument; they determined, boldly among themselves they were the rising stars, they were the lights of the chambers, where the utmost importance was the form of upward motion, the perfection they encumbered, to examine and express themselves in all their grandeur as if they had been conscious of it. It was a sunburst of sound, breaking out of hiding and spreading over the whole room. It had the freedom of release and the tension of purpose. It swept space clean. There was a feeling from which the music had escaped, and spoke in laughing astonishment at the discovery that there was no ugliness or pain, and there never had had to be. It was the sound of a great pianoforte.

Style score difference towards Austen: + 0.03932239

Classifier naturalness score difference towards the original: + 0.05681665

D Content example: Jane Austen, "Pride and Prejudice"

Elizabeth listened in silence, but was not convinced; their behaviour at the assembly had not been calculated to please in general; and with more quickness of observation and less pliancy of temper than her sister,

```
and with a judgement too unassailed by any attention to herself, she
was very little disposed to approve them. They were in fact very fine
ladies; not deficient in good humour when they were pleased, nor in the
power of making themselves agreeable when they chose it, but proud and
conceited. They were rather handsome, had been educated in one of the
first private seminaries in town, had a fortune of twenty thousand
pounds, were in the habit of spending more than they ought, and of
associating with people of rank, and were therefore in every respect
entitled to think well of themselves, and meanly of others. They were of
a respectable family in the north of England; a circumstance more deeply
impressed on their memories than that their brother's fortune and their
own had been acquired by trade.
```

Style score w.r.t. Nabokov: 0.0057041

Classifier naturalness score: 0.9612681

D.1 Machine Generated: Nabokov style transfer

Elizabeth listened in silence, but then understood that this was a deliberate lie; their behaviour at the assembly had not been calculated to please in general; and with more quickness of observation and less pliancy of temper than her sister, she discerned in her a deceptive, though quite reasonable, form of reflection, and not only because she was concealing her ineptitude to keep up with the jol. They were in fact very fine ladies; not deficient in good humour when they were pleased, but parched; not easily swayed by carnal particularities; not indifferent minds when it came to but proud and conceited. Their meeting with everybody was very warm and well-wishing; and their smiling faces, had been educated in one of the first private seminaries in town, had a fortune of twenty thousand pounds, were in the habit of spending more than they ought, and of associating with people of rank; and were therefore in every respect entitled to think well of themselves, and meanly of others. The greater part of their city consisted of solidly cottoned Estotians; a circumstance more deeply impressed on their memories than that their brother's fortune and their own had been acquired by trade.

Style score difference towards Nabokov: + 0.00143078

Classifier naturalness score difference towards the original: - 0.0104554

E Content example: Alexander Dumas, "The Vicomte de Bragelonne"

```
In the meanwhile, Monsieur continued his route with an air at once so
melancholy and so majestic, that he certainly would have attracted the
attention of spectators, if spectators there had been; but the good
citizens of Blois could not pardon Monsieur for having chosen their gay
city for an abode in which to indulge melancholy at his ease, and as
often as they caught a glimpse of the illustrious ennuye, they stole
away gaping, or drew back their heads into the interior of their
dwellings, to escape the soporific influence of that long pale face, of
those watery eyes, and that languid address; so that the worthy prince
was almost certain to find the streets deserted whenever he chanced to
pass through them.
```

Style score w.r.t. Nabokov: 0.00651407

Classifier naturalness score: 0.8818566

E.1 Machine Generated: Nabokov style transfer

In the meanwhile, Monsieur continued his route with an air at once ardent and melancholy, and I suppose it gave him a thrill of recognition that I doubt it would be in the course of the of spectators; but the good citizens of Blois could not pardon Monsieur for having chosen their gay city for an abode in which to indulge melancholy and oppress him, and as often as they caught a glimpse of his manipulation, they stole away gaping, or drew back their heads into the interior deals of citizens, to escape the soporific influence of that long-dressed hoodlum with purplish wrinkles and flaming cheekbones, and blue eyes, and that languid address; so that the worthy prince was almost certain to find the streets deserted whenever he chanced to glance at them.

Style score difference towards Nabokov: + 0.00031625

Classifier naturalness score difference towards the original: + 0.1134084